Cognition 142 (2015) 191–204

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

The kind of group you want to belong to: Effects of group structure on group accuracy

Martin L. Jönsson [a,*], Ulrike Hahn [b], Erik J. Olsson [a]

[a] Department of Philosophy, Lund University, Sweden
[b] Department of Psychological Sciences, Birkbeck University of London, United Kingdom

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 11 April 2015
Accepted 18 April 2015

Keywords:
Wisdom of crowds
Group structure
Group accuracy
Information flow
Social networks
Group learning

ABSTRACT

There has been much interest in group judgment and the so-called 'wisdom of crowds'. In many real world contexts, members of groups not only share a dependence on external sources of information, but they also communicate with one another, thus introducing correlations among their responses that can diminish collective accuracy. This has long been known, but it has—to date—not been examined to what extent different kinds of communication networks may give rise to systematically different effects on accuracy. We argue that equations that relate group accuracy, individual accuracy, and group diversity (see Hogarth, 1978; Page, 2007) are useful theoretical tools for understanding group performance in the context of research on group structure. In particular, these equations may serve to identify the kind of group structures that improve individual accuracy without thereby excessively diminishing diversity so that the net positive effect is an improvement even on the level of collective accuracy. Two experiments are reported where two structures (the complete network and a small world network) are investigated from this perspective. It is demonstrated that the more constrained network (the small world network) outperforms the network with a free flow of information.

1. Introduction

Interaction with others in different social groups is an essential part of the human condition. As members of a jury we deliberate with fellow jurors in order to arrive at an appropriate verdict, as members of a legislative body we interact with others to create and repeal laws, and as members of research groups we pool our resources so that we jointly can perform better than we can do individually.

We frequently trust the verdicts and estimates of our groups, even in cases where they are in conflict with our own. The well-foundedness of this trust has been the subject of much research in social psychology. Early on, Galton (1907) famously compared the accuracy of a group with that of its members in guessing the weight of an ox during a stock and poultry exhibition. During subsequent decades, social psychologists carried on in this tradition by comparing groups with their members on a variety of tasks, from the estimation of room temperature, to the judgment of children's intelligence from photographs, to the solution of mathematical problems (see, e.g. Knight, 1921; Shaw, 1932, for extensive reviews see Gigone & Hastie, 1997; Hill, 1982; Lorge & Brenner, 1958). The bottom line of much of this was that, on the one hand, results could not really be made sense of without formal statistical tools, and, on the other, that once these were properly utilized, much of these earlier results seemed trivial.

In the words of Gigone and Hastie (1997).

> Statistical combinations of judgments have long been known to cancel out unsystematic judgment error

* Corresponding author.

http://dx.doi.org/10.1016/j.cognition.2015.04.013
0010-0277/

(Hogarth, 1977). The standard error of the mean of several judgments is smaller than the standard deviation of the judgments themselves; groups almost inevitably outperform their members simply by averaging those members' judgments. Such accuracy gains can hardly be attributed to anything special about the group judgment process; the group need not meet at all.

Gigone and Hastie (1997: 159)

So the real question of group research must be the extent to which the group is better than the statistical aggregate (Gigone & Hastie, 1997). Or, to put this differently, what is it that the group adds?

One way of approaching this question is to manipulate communication channels within a group and examine attendant effects. Experimental manipulation of the information participants receive from others allows inference about the extent to which they use that information. It thus provides a methodological window into how people go about combining what they believe with information they receive from others.

This question seems at least as relevant now as it did in the early days of small group research, because it has become ever more apparent that our beliefs and opinions are determined not merely by our own observations, but, to an arguably even greater extent, by the evidence we receive through the testimony of others (see e.g., Coady, 1992). Consequently, there is only so much one can study about human learning, judgment and decision making without taking into account the social dimension of belief formation (see also, Goldstone & Gureckis, 2009).

This in turn suggests a subtle shift in emphasis concerning the kinds of groups and tasks that are of interest and what aspects of group influence and performance seem most worthy of examination. Much of the past research on groups (as surveyed in the reviews of Gigone & Hastie, 1997; Hill, 1982; Lorge & Brenner, 1958) has focussed on the quality of the group response itself, and this is also the central theme in the recent revival of this tradition of research under the header of 'wisdom of crowds' (Hertwig, 2012; Herzog & Hertwig, 2009; Surowiecki, 2004). However, it is at least as interesting and important to ask what the group does for the individual, and how this develops, that is, to ask not just how group performance compares to individual performance but to ask how both individual and group performance are changed by group communication.

It is here that useful links can be formed with the burgeoning literature on networks, in particular social networks (for an introduction see e.g., Jackson, 2010). Patterns of communication between individuals in groups give rise to network structure (see also, Goldstone, Roberts, & Gureckis, 2008; McGrath, Arrow, & Berdahl, 2000): depending on context, all members may be exchanging views and listening to one another freely; alternatively, only some members may be communicating directly with one another. Finally, even where all individuals hear all information being exchanged, selective attention and weighting (see e.g., Friedkin & Johnsen, 1999) of others' information (determined, for example, by perceived competence) imposes an effective network structure to the communication that diverges from the surface level whereby everyone is communicating with everybody else.

As just indicated, experimental manipulation of the structure of communication may provide insight into what it is that being part of a group is adding. At the same time, it raises interesting questions of its own concerning the extent to which different types of communication networks may systematically differ in their impact on our beliefs (on the general benefits of taking a network perspective to traditional group research see also, Katz, Lazer, Arrow, & Contractor, 2004).

To this end, we present two experimental studies manipulating the communication structure within a group and examining its impact on the accuracy of participants' beliefs. To sidestep some of the pitfalls of the early work on group accuracy, our analysis is informed by two equations that relate group validity, individual validity and group diversity. These equations demonstrate—for two different ways of aggregating opinion and two different ways of understanding accuracy—the conditions under which the group will outperform its average individual member by mathematical necessity. First, work by Ghiselli (1964, chap. 7) and Hogarth (1978) points out that if the validity of a sequence of estimates is understood in terms of the correlation between it and the true values, the validity of the group estimate can be shown to always exceed the average validity of the answers of the group members as long as the members are not perfectly correlated with each other and error is unbiased.

More precisely, if we let n be the number of group members, $s_{x_i}, s_{\bar{x}}, s_t$ be sequences of the estimates of group member i, mean estimates, and true values respectively, and $\rho_{x,y}$ be the correlation between two sequences x and y, Hogarth's equation states that[1]

$$\rho_{s_t s_{\bar{x}}} = \frac{\sqrt{n}\frac{\sum_{i=1}^{n}\rho_{s_t s_{x_i}}}{n}}{\sqrt{1+(n-1)\frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho_{s_{x_i} s_{x_j}}}{n}}} \quad (1)$$

The limiting case where $n = \infty$ is captured by following equation:

[1] It should be noted that understanding validity in terms of a correlation results in a fairly coarse-grained concept of validity. For instance, assume that Bob and Sue have answered in the following way:

	Bob	Sue	Correct
Question 1	13	5	5
Question 2	15	7	7
Question 3	11	3	3
Question 4	13	5	5

On the correlational understanding of validity, Bob's and Sue's answers are, counterintuitively, equally valid (both answers are perfectly correlated with the correct answer). This might be what Hogarth is after when he remarks that his results only hold in circumstances where 'the judgmental task consists of rank ordering alternatives—that is the *level* of judgment is not important.' (Hogarth, 1978: 41, emphasis in original). Nonetheless, even when the exact values are important for a correct answer, the correlation between a sequence of answers and the correct answers gives us an indication of how good the answers are; answers that are very poorly correlated with the correct answers cannot be correct.

$$\lim_{n\to\infty}\rho_{s_t s_{\bar{x}}} = \frac{\frac{\sum_{i=1}^{n}\rho_{s_t s_{x_i}}}{n}}{\sqrt{\frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\rho_{s_{x_i} s_{x_j}}}{n}}} \quad (2)$$

More generally, the Diversity Prediction Theorem (Page, 2007) indicates that the collective error of a group (measured as the squared deviation between true value and group mean) must equal the average individual error minus the group diversity (measured as the variance around the mean estimate). It follows directly from the theorem that as long as there is diversity in the group, the collective error must always be lower than the average individual error.

More precisely, if we let n be the number of group members, x_i be the estimate of the group member i, $\bar{x}$ be the mean estimate, and t be the true value of whatever is being estimated, the Diversity Prediction Theorem states that

$$(\bar{x}-t)^2 = \frac{\sum_{i=1}^{n}(x_i-t)^2}{n} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \quad (3)$$

Both Hogarth's equation and the Diversity Prediction Theorem thus make clear that group diversity decreases collective error: more diverse groups will give rise to more accurate collective judgment.

Although these equations are of general importance in research on group interaction, they have not hitherto informed research on group structure. As a corollary to our main findings we hope to demonstrate the usefulness of these equations not only in understanding data on group performance, but also on suggesting interesting empirical predictions and navigating theoretical arguments about the conditions for successful group interaction.[2]

2. A dim view of social interaction

The fact that the group will perform at least as well as its average individual does not mean that the group always performs well; if the average individual validity is very low then so is the group validity. The truth of the equations described above is thus compatible with the view held by some researchers, that group interaction is often not very beneficial. Lorenz, Rauhut, Schweitzer, and Helbing (2011), for instance, adopt this view (see also King, Cheng, Starke, & Myatt, 2012). In support of their view, Lorenz et al. report an experiment where participants were divided into groups of twelve people and asked to answer questions under three different conditions. Under one of these conditions, a participant first answered a question on her own, was then given information about what the other members of the group had answered, and was then asked to answer the same question again in light of the new information. This procedure was repeated three more times. Even though it might be expected that the participants had excellent opportunity to improve their guesses under this condition, Lorenz et al. did not find any significant improvement in collective error at any point during the five rounds, but did find that the group had became significantly less diverse. They concluded that the group is actually worse off by the end of the five rounds than when it starts out, in part due to this loss of diversity without corresponding decrease in collective error.

It is possibly premature for at least two reasons to adopt Lorenz et al.'s somewhat pessimistic position on the basis of their experiment. First, as Farrell (2011) remarks, Lorenz et al.'s data actually show that the participants improve individually in the relevant condition, even though there is no decrease in collective error. In fact, it follows directly from the Diversity Prediction Theorem that if we have stable collective error and a decrease in diversity, we must also have a decrease in the average individual error. So Lorenz et al.'s experiment did demonstrate that social interaction is beneficial at the individual level. Second, in another of the three conditions ("the aggregate information-condition") in Lorenz et al.'s experiment, participants were given the group's mean answer as feedback after each individual guess, and in this condition a significant decrease in collective error did occur. Our analysis of their data also shows that the average individual error decreased significantly in this condition. In line with Lorenz et al.'s own procedure, we concluded this by comparing the logarithms of the participants estimates divided by the true values. In the aggregate information condition there were 24 groups (six questions answered by four groups each). The average individual error in the first round was compared with the average individual error in each of the subsequent rounds. A repeated measures two-tailed t-test revealed all differences to be significant ($p < 0.01$). The trends for individual error, collective error and diversity in the aggregate information condition can be seen in Fig. 1.

This means that in Lorenz et al.'s experiment, one condition gave rise to significant decrease in both collective and average individual error. So in the two conditions where participants were given feedback it was only in the full information condition that the collective error did not decrease significantly (see also Yaniv & Milyavsky, 2007).

An appropriate conclusion to draw from Lorenz et al.'s data is thus that there are conditions in which group performance improves both at the collective level and at the individual level, but in some groups there is only the latter kind of improvement. The Diversity Prediction Theorem is useful here since it straightforwardly suggest an analysis of the found contrast: it follows from the theorem that if the diversity of a group decreases about as much as the average individual error, there is no decrease in collective error. The full information condition in Lorenz et al.'s experiment, where every participant can see every other participant, follows this pattern, and the theorem can thus explain the absence of a decrease in collective error. Despite the fact that there is decrease in individual error, the decrease in diversity is so considerable that it negates the individual improvement.

This analysis raises the very general question of what properties of groups promote decreases in individual error without giving rise to excessive decreases in diversity. If we can identify these properties, we can identify the

[2] For two interesting applications of the Diversity Prediction Theorem to the 'collective' constituted by a single person at different times, see Rauhut and Lorenz (2010) and Vul and Pashler (2008).

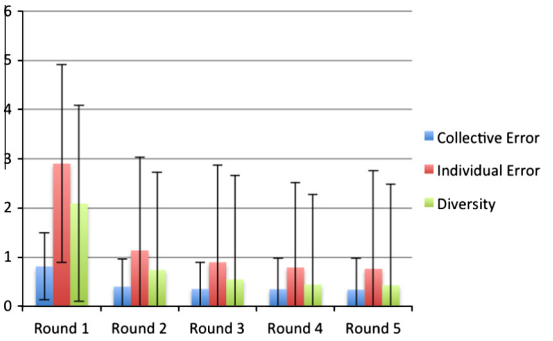

Fig. 1. Reanalysis of data from Lorenz et al. (2011): Trends for individual error, collective error and diversity in the aggregate information condition. Error bars correspond to ±1 SD.

groups that are ideal for social interaction in the sense that they both decrease individual and collective error. In this paper we will contribute to answering this question in the context of the research in group structure. As a starting point, we will use structures that research on networks has identified as having independent theoretical interest. We next provide an overview of research relevant to the question of network structure and group performance.

3. Groups with a restricted flow of information

Given evidence from simulation for the importance of network structure for contagion and diffusion (see e.g., Kretzschmar & Morris, 1996; Lazer & Friedman, 2007; Watts, 1999; see Jackson, 2010; for an introduction), including such processes as information dissemination (e.g., Doer, Fouz, & Friedrich, 2012), it seems reasonable to assume that network structure would be found to influence actual judgments by real people. However, there has so far been no experimental study of the impact of network structure on estimation tasks such as those that have been the focus of research on group influence and 'wisdom of crowds'.

There are, however, two decades worth of experimental studies that manipulated network structure in the context of collective problem solving, finding evidence for a causal role of network topology (for a review see e.g., Shaw, 1964; for a brief summary see also, Levine & Moreland, 2012; for recent work resuming aspects of this tradition, see Kearns, 2006). In the classical work in this tradition, founded by Leavitt (1951) and Bavelas (1950), participants were assigned to laboratory-based, *ad hoc* groups within the context of an apparatus that constrained communication channels. For example, participants were seated such that they were screened off from each other by dividers and could communicate only via written messages passed through slots in these dividers. The task participants faced were simple collective problem solving tasks which required the combination of information held across participants for their solution. For example, each participant might receive a hand of five cards each displaying a different symbol, and the collective task was to identify the one symbol shared by all. The key experimental manipulation in these studies was the network structure of the communication channels: for example, chain-like arrangements were contrasted with wheel (circle), and star-like configurations with a central actor. Typical dependent variables of study were not only performance indicators such as time taken to complete the task and error rates, but also degree of satisfaction with group membership and measures of leadership within the group. In general, the theoretical emphasis within this tradition was on centralization and the role of putative leaders on group performance (see also Freeman, Roeder, & Mulholland, 1979).

While there are indeed many real-life problems that require collective problem-solving in this sense, that is, the combination of skills or information held uniquely by different individuals in order to achieve a particular goal as a group, that is, a single problem-solving unit, this by no means delimits the role of social influence and groups in everyday life. In particular, in the context of belief and opinion formation, our goals are, in first instance, individualistic: it is the accuracy/informedness of our own beliefs and opinions that we care about, not how those opinions might contribute to 'collective knowledge' of the group. In forming those beliefs, we may choose to consult the views of others, but it is not necessary that we do so. Needless to say, as social beings it is often also of great importance to us that we influence the beliefs and opinions of others, and it may on many occasions be of importance for us to do so in order to achieve our goals, which may require some form of collective action. However, it is our contention that it is characteristic of human beings that they acquire information even where there are no immediate actions apparent that may follow from that knowledge, and that in those contexts too, social influence is pervasive.

In keeping with this, we wished to study the impact of others on a purely knowledge based estimation task which, though it exposed people to others' opinions did not require them to consider those opinions in any way. This context matches that of much early 20th century small group research (as discussed in the Introduction above and in, e.g., Gigone and Hastie (1997)) that was focussed on accuracy, but that work did not consider network structure. The previous work most closely related to our aims is the study by Mason, Jones, and Goldstone (2008) where different network structures were tested in terms of how well they supported searches in multimodal problem spaces. Among the networks tested was *the complete network*, i.e. the network where everyone can see everyone else, *a small-world network*, and *a random network*, all three of which are network structures of independent theoretical interest (more on this below) and examples of which are given in Fig. 2.

In one of their experiments, Mason et al. (2008) found that the participants in a small-world network were actually faster than those in a fully connected network in finding a global maximum, that is, the group where the information channels between its members were restricted performed better. Although Mason et al.'s explanation of their result does not carry over to the estimation tasks of present concern, the structures they used are

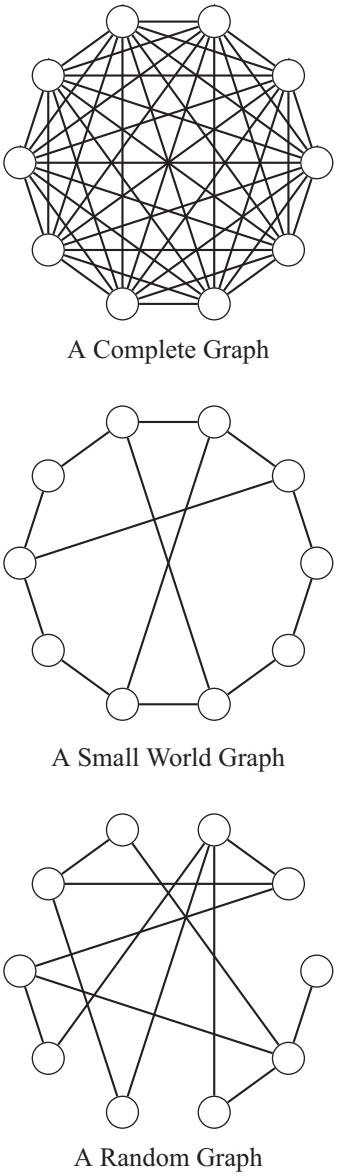

Fig. 2. Network types examined by both Mason et al. (2008) and the present study.

interesting from our perspective.[3] These structures might be able to capture the contrast between excessively diversity decreasing networks, and networks that reduce diversity to a more accuracy enhancing degree.

The Diversity Prediction Theorem entails that collective error decreases only if the average individual distance to the truth decreases faster than the average individual distance to the mean. Too much communication may reduce diversity too much to allow this to happen. Like Mason et al. we thus examined a complete network as a baseline with which to check whether 'more information' is always better, given that people are free to ignore parts of it. We also examined a random network (a so-called Erdos–Renyi random network as is formed by generating links independently and randomly with a particular probability p, Erdos & Rényi, 1959), because these form a common baseline of comparison in the network literature. Finally, we examined a small world network (see Watts, 1999; Watts & Strogatz, 1998), a type of network that is characterized by short paths between nodes and higher clustering than observed in random graphs, and which seems to capture not just of a wide variety of biological networks, but also, most importantly, social networks.

4. Experiment 1

In the first experiment we studied a percentage estimation task in medium sized groups ($N \approx 9$) in order to determine whether networks with more limited connectivity than the full network would result in decreases in average individual error without excessive decreases in diversity so that the net result would be a significant decrease in collective error. The experiment also tested the accuracy of Hogarth equation in the context of interacting groups. Unlike the Diversity Prediction Theorem—which is completely general—Hogarth's equation assumes unbiased error, and is assumed to apply necessarily only to groups of non-interacting members, that is, so called 'staticized groups' (see Einhorn, Hogarth, & Klempner, 1977; Hogarth, 1978). It is thus empirically interesting how well the equation will fare in circumstances where participants have some degree of interaction and unbiased error cannot be guaranteed.

5. Method

5.1. Participants

28 undergraduate students (18 male and 10 female) at Lund University were paid for their participation. In addition to a flat reward for participation (100 SEK), participants were incentivized to take the task seriously by a reward corresponding to three times the participation reward (300 SEK) to the person in each group with the most accurate answers. All participants gave written and informed consent to the experimental procedure.

5.2. Procedure and materials

Each participant signed up for one out of three different dates, and was tested together with other participants that signed up for the same date. Each of the resulting groups (containing 10, 10 and 8 participants respectively) was tested separately, but the testing conditions, procedure, materials and instructions were the same across the three groups.

[3] Mason et al.'s explanation echoes March (1991) who argued that homogenous groups do too much exploiting of known solutions and too little exploring to find new solutions (see Lazer & Friedman, 2007 for a similar simulation result, and a similar explanation). This kind of explanation is only applicable in situations where people are given objective feedback about the success of their estimate or guess, because the notion of exploration only makes sense in this context (consider for instance a completely sensory deprived cartographer).

During the experiment, each member of a group was seated at a computer and was given two sheets of paper with instructions. When everyone in a group stated that they had understood the instructions, a server program was run which sent out questions to all participants.[4] There were thirty target questions in total, and an initial warm-up question. Each question was asked five consecutive times over the course of five consecutive rounds. During the first round, each participant had to answer the question without any knowledge about what the other participants in the room thought the answer was. For the next four rounds, each participant received information about what other participants had answered on the previous round.

Each of the questions was associated with one of the three network types described above, so that there were 10 questions associated with each network-type. With the start of the second round of each question, a network of the appropriate type was randomly generated such that it contained one node corresponding to each participant. Participants could then, on the subsequent rounds, see the answers of the participants corresponding to the nodes they were immediately connected to. The participants did not know anything about the overall structure of the network but they were informed that other participants might be able to see people they themselves could not see. Furthermore, they were told that every score mattered equally. The participants were only told the correct answers after the experiment was over. The winner of each group was determined by calculating the correlation of each participant's sequence of scores with the correct scores.

We adopted Mason et al.'s (2008) strategy for generating networks. The complete network was created by generating $n(n-1)/2$ (where n is the number of nodes) edges so that every node was adjacent to every other node. The small-world network was created by first connecting all nodes in a ring, and then randomly combining 30% of the nodes with another node at least three nodes apart from it (following the ring path). Finally, the random network was generated by creating a number of edges equal to 1.3 times the number of nodes under the constraint that a path should exist between every two nodes. Both the small-world network and the random network thus have the same average degree ($2.6n/n$) and the same link density ($2.6n/(n*(n-1))$).

We devised a set of 31 questions for use in the experiment from reports by Statistics Sweden ('Statistiska Centralbyrån', a well established administrative agency which supplies the Swedish government and other organizations with statistics) on, for instance, Swedish demographics, agriculture and geography (all questions are listed in Appendix A). Except for the warm-up question, the questions were presented in a random order. The randomization was the same across the three groups.

We had to choose whether to make network-type a within or a between participant variable. Clearly, question difficulty is easier to match if the same question are associated with each network structure, but this is possible only in a between participant design. However, not just question difficulty will vary (and how difficult different individuals will find them), but arguably also how much people pay attention to the opinions of others may be subject to individual differences, and this latter factor can be controlled only by making network structure a within participant variable (on the importance of within-group manipulations in group research see also McGrath et al., 2000). Consequently, we opted for a within-subjects design whereby each participant experienced all network structures. This approach also had the added benefit of allowing us to test more networks per participant (counterbalancing the questions across the three network-types would have required three times as many participants, or alternately, testing fewer networks). We thus estimated the difficulty of the questions from a pilot study, calculating the average distance (in percentage points) between the participants' answers to the questions without feedback and the correct answer. The questions were then assigned to networks so that the average difficulty of the questions (as estimated by a pilot-test) and the spread in correct answers from 0–100%were roughly the same for the three networks.

6. Results

The data from the three groups were pooled so that measures of collective accuracy would reflect the judgments of equal numbers of participants. From the pooled data, the diversity (variance), average individual error (mean squared error across individuals), and collective error (squared deviation of the mean answer from the true value) were calculated for each round for all networks. There are two ways one could calculate these three quantities: either by directly calculating them from the corresponding 280 data points per network type (28 participants times 10 questions), or by calculating each quantity first for each group and then taking the mean of those calculations (3 groups times 10 questions). In the context of this experiment, there is little difference between these strategies (other than that it slightly boosts the influence of the smallest, 8 person group). However, we used a between participant design in Exp. 2 below, and there the second strategy seems preferable as it ensures the integrity of the counter-balancing of questions and networks. Because we wanted to keep the analysis of both studies comparable, we used the second method for Exp. 1 also.

The results of these calculations can be seen in Table 1. As is entailed by the Diversity Prediction Theorem, the group outperforms the average individual across all network types and rounds, and the difference between the collective error and the average individual error corresponds to the group diversity.

In general, both individual error and collective error decreased across consecutive rounds for all network types, as did the diversity of the raters. However, only some of these decreases were statistically significant. For each network type, the diversity, individual error, and collective error in the first round, were compared to the

[4] The program was written in NetLogo by the first author by adapting for present purposes a program kindly provided by Rob Goldstone, which is based on routines from Netlogo's Hubnet sample code.

Table 1
Experiment 1: Diversity and error, means, by network type (complete graph, small world network and random network), and judgment round on a given question.

Network	Rnd	Diversity	Ind. error	Col. error
Complete	1	190.0	314.4	124.5
Complete	2	144.0	253.1	109.1
Complete	3	116.4	216.2	99.8
Complete	4	86.7	181.7	95.0
Complete	5	76.0	164.0	88.0
Random	1	256.0	412.7	156.7
Random	2	172.2	318.9	146.7
Random	3	147.9	289.5	141.6
Random	4	143.1	283.7	140.6
Random	5	138.2	269.4	131.2
Small World	1	169.4	464.7	295.3
Small World	2	112.1	419.8	307.7
Small World	3	93.0	385.2	292.2
Small World	4	85.1	371.1	286.0
Small World	5	76.6	366.4	289.8

corresponding quantities in every subsequent round. Paired two-tailed *t*-tests revealed significant decreases in diversity and average individual error that were maintained across subsequent rounds for all network-types (in all but one case $p < .01$, and in most cases $p < .001$). Comparisons were done across questions and groups (with resulting $df = 29$). With respect to collective error, decreases were not significant at the level of the individual network as can be seen in Table 2 which reports the corresponding tests for collective error. This, however, given the trends in the data, reflects a lack of statistical power more than anything else, as is apparent from the fact that, collectively, across all network types, round 5 error was significantly lower than initial (round 1) error, $t(89) = 2.13, p = .036$.

The data were also analyzed from a correlational perspective in order to gauge the accuracy of Hogarth's Eq. (2), in the circumstances of this experiment. Across the three groups, and across the three network types, the equation predicted the validity of the mean answer (i.e., its correlation across questions with the true answer) with a high degree of accuracy, with a mean prediction error below .0065. The overall degree of correlation between the observed validity of the mean answer and the predicted validity was very high (≈.99). The accuracy of the equation was comparable across the three network types. As is entailed by the equation, the mean validity outperformed the average individual validity across rounds and network-structures.

Finally, we wished to see whether there were consequences of the differences in network structure. A direct comparison of accuracy (individual or collective) is not meaningful for these data as the questions (despite our attempts to balance them) varied in difficulty across network types. The mean individual errors were 314, 413 and 465 respectively for the complete network, the random network and the small world network. This difference between these scores suggests that the questions assigned to the three network types were of uneven difficulty, and that the two constrained networks were associated with harder questions than the unconstrained network. Two

Table 2
Experiment 1: Comparison between round 1 collective error and collective error at each subsequent round for each network type, numbers displayed are *p*-values for two-tailed *t*-tests.

Network	Round 2	Round 3	Round 4	Round 5
Complete	.087	.232	.107	.076
Random	.563	.436	.427	.223
Small World	.403	.846	.560	.699

two-tailed *t*-tests revealed that this difference was marginally significant between the complete network and the small world network ($t(29) = -1.83, p = .072$) but not significant between the complete network and the random network ($t(29) = -1.18, p = .24$).

One can, however, ask to what extent network structure led to correlations between raters, in particular, to what extent it led to correlations that are not simply reflections of the increase in accuracy. Degree of accuracy constrains the inter-rater correlation across participants, in that greater accuracy necessarily means greater inter-rater correlation, even in the case where their estimates are entirely independent: in the limit where participants' estimates are perfectly correlated with the data, they must also be perfectly correlated with each other.

However, one can ask to what extent participants' estimates show correlations above and beyond the level mandated by their level of accuracy. Specifically, we devised a measure of 'excess correlation' that factored out the absolute level of accuracy and could thus be compared across network structures, indicating the extent to which a network was associated with correlations between participants that were not validity inducing. Given that the answers of two participants *a* and *b* are both correlated with the true values to a certain degree, there is an interval in which their degree of correlation *with each other* must lie. The maximum of this interval is the perfect correlation that would arise if *a* and *b* always gave identical answers. The minimum of this interval is the lowest degree of correlation that could obtain between them without decreasing their accuracy, that is, the correlations of either (or both) of their responses with the true answers. All correlation above this minimum, lowest degree is 'excess correlation': that is, correlation not required for the persons to have their respective validities. In other words, excess correlation measures the extent to which pairs of raters are more correlated than they would need to be, given their respective levels of accuracy.

If the degree of actually observed excess correlation is normalized (that is, considered as a proportion of the remaining interval spanned by the minimum necessary correlation and a perfect correlation of 1), it provides a measure of correlation excess that can be compared across different absolute levels of accuracy. So by calculating the pairwise minimum correlation for all participants (given their validities) we derived a measure of the excess correlation in each condition (i.e., $1 - (1 - \rho_{a,b}/1 - \rho_{a,b}^{Min})$) that ranges from 0 (no excess correlation) to 1 (maximum excess correlation). On this metric, the complete network led to the greatest excess correlation (+.1417) closely followed by the random network (+.1416), with the lowest

degree of excess correlation found in the small world network (+.1071). This suggests that a complete network, relative to a small-world network, may increase inter-dependence between raters beyond the extent to which it is useful for improving accuracy. We pursued this question further in Experiment 2.

7. Discussion

Previous research in the context of advice taking has suggested that participants overweight their own judgement and underweight the responses of others relative to what would be optimal, and that participants benefit only from fairly few opinions, becoming less sensitive to the recommendations or judgments of others as these increase beyond just a handful of opinions (see Yaniv & Milyavsky, 2007). In line with this, participants in Exp. 1 did not attach equal weight to their own opinions and those of others. This is apparent from the fact that participants did not simply converge on the mean of all answers in round 2 of the complete network condition. Equally weighting all visible responses, including one's own, would be equivalent to taking the mean of those responses to be one's revised prediction, with the consequence that individual error would come to equal collective error on round 2. At the same time, the fact that individual error remained higher (even on the last round) than the collective error in the initial round of independent judgments (round 1) indicates that participants' failure to do so came at an accuracy cost. To this extent, the findings of Yaniv and Milyavsky (2007), who showed participants only one round of other responses, are confirmed in our study. However, it is surprising from the perspective of that research how much information from others our participants were actually willing to take in. In particular, diversity decreased round on round (see Table 1) indicating not only that (at least some) participants continued to revise their opinions round after round, but that they did so in a way that was sensitive to the responses of others.

Our results also illustrate clearly how the formal tools of the Diversity Prediction Theorem and Hogarth's equation can be used to understand the relationship between collective accuracy, individual accuracy and inter-rater dependence. Collective accuracy, individual accuracy and the variability of the individual raters necessarily trade off, and, all other things being equal, collective error is negatively affected by decreases in variability brought about by communication.

However, contrary to the slightly negative view expressed by Lorenz et al. (2011), average individual error was significantly reduced in every round following the first one, even though diversity decreased as a result of information exchange. This is in line with the observations made by Farrell (2011) regarding the (unreported) individual-level improvements in Lorenz et al.'s data. Finally, in even starker contrast to the sceptical view on social interaction of Lorenz et al. (2011) there was a significant decrease in collective error by round 5 in the overall data.

From a correlational perspective, it can be concluded that Hogarth's equation can be used to make remarkably accurate predictions of the validity of the mean answer based on the average individual validity and the average interpersonal correlation in the group, even in the case of explicitly interacting groups. Such interactions may well serve to accentuate any biases present in participants' responses, but the equation may be quite robust to violations of its assumptions (as has also been found in related work on judgment aggregation by Wallsten, Budescu, Erev, & Diederich, 1997).

Furthermore, our measure of excess correlation showed, as predicted, that the complete network increased more in excess correlation than the more constrained networks, with the small world network increasing the least of the three networks. However, the absolute degree of excess correlation observed may also be influenced by the difficulty of questions, and the difference in excess correlation between the complete networks and more constrained networks might thus be even more pronounced when the difficulty of questions is the same across the networks. Though our measure of excess correlation takes into account variations in the degree of accuracy, and hence question difficulty, this measure could not take into account any difference in participant strategy that emerge as a function of difficulty. It seems entirely possible that the difficulty of questions influences how much people benefit from social interaction. At one extreme, if no-one in the group has any information about the correct answer, interacting socially will generate no benefits. Thus to further pinpoint the relative merits of different network structures, the difficulty of questions needs to be more closely matched. This was done in Experiment 2.

8. Experiment 2

Experiment 2 was very similar to Experiment 1 but used the alternative strategy of making network-type a between participant variable. It thus controlled directly for differences in the difficulty of questions across different network structures, by having, across groups, all questions be associated with all networks. It also extended the number of rounds per question from five to eight in order to make sure that trends in average individual error, diversity or collective error were not interrupted prematurely. In order to accommodate these changes without the experiment becoming excessively long, only two network-types were tested, the complete network and the small-world network, which is the most interesting of the two restricted networks examined in the last study, given that real world social networks typically have small world structure (Watts, 1999).

9. Method

9.1. Participants

38 undergraduate students (15 male 23 female) at Lund University were paid for their participation. The participants were rewarded and incentivized in the same was as in Experiment 1. All participants gave written and informed consent to the experimental procedure.

Table 3
Experiment 2: Diversity and error; means.

Network	Rnd	Diversity	Ind. error	Col. error
Complete	1	237.7	401.6	163.9
Complete	2	121.7	249.2	127.5
Complete	3	102.1	227.9	125.8
Complete	4	112.0	238.9	126.9
Complete	5	110.2	234.4	124.2
Complete	6	109.6	246.3	136.7
Complete	7	105.2	233.5	128.3
Complete	8	102.5	230.9	128.4
Small World	1	279.7	411.9	132.3
Small World	2	177.9	263.7	85.8
Small World	3	167.5	248.6	81.2
Small World	4	140.2	209.3	69.0
Small World	5	136.6	204.3	67.7
Small World	6	137.3	207.3	70.0
Small World	7	130.2	194.6	64.4
Small World	8	138.4	213.6	75.2

Table 4
Experiment 2: Collective error: comparison of subsequent rounds to round 1 error, *p*-values.

Network	R2	R3	R4	R5	R6	R7	R8
Complete	.070	.111	.172	.169	.384	.230	.248
Small World	.008	.029	.007	.006	.009	.005	.008

9.2. Procedure and materials

Each participant signed up for one out of four different dates, and was tested together with the other participants that signed up for the same date. Each of resulting groups (containing 9, 9, 7 and 13 participants respectively) was tested separately. The testing conditions, procedure, and instructions were the same across the four groups.

The testing arrangement was very similar to that used in Experiment 1 except for the fact that each participant answered each question eight times (rather than five) and the fact that there were only twenty target questions in total (a subset of the questions used in Experiment 1). Moreover, each of the questions was associated with one of only two network types: the complete network and the small world network. There were 10 questions associated with each network-type. The question-network association was counter-balanced across groups so that the set of questions that groups 1 and 2 answered from the perspective of the complete network were answered from the small-world network by groups 3 and 4 and vice versa.

10. Results

Four scores were eliminated as likely errors before data-analysis was conducted. Three of these were zero-answers that likely resulted from the participant accidentally clicking submit before choosing his or her estimate (which was done on a sliding bar next to the submit-button), and one was a very large number in a sequence of identical low numbers which was also likely to be due to a mis-click.

The data from the four groups were pooled and the diversity, average individual error, and collective error was calculated for each round for both networks, following the same procedure outlined for Exp. 1 above. The results can be seen in Table 3.

As is entailed by the Diversity Prediction Theorem, the group outperforms the average individual across network types and rounds, and the difference between the collective error and the average individual error corresponds to the group diversity.

For each network type, the diversity, individual error, and collective error in the first round were compared to the corresponding quantities from every subsequent round. Paired two-tailed *t*-tests revealed significant decreases in diversity and average individual error across all rounds and both network-types (in all cases $p < .005$). Comparisons were done across questions and groups (with resulting $df = 39$). Table 4 reports the corresponding tests for collective error. The trends for the three quantitates across all eight rounds (comparing to round 1) can be seen in Figs. 3–5.[5] As can be seen there is a stable reduction in collective error only for the small world network condition.

The small world network (but not the complete network) also continued to show improvements for later rounds: both individual and collective error dropped significantly below that of round 2 from round 4 onwards [6], staying significantly below the round 2 levels, barring a blip in the final round for collective error (see Fig. 3).

The pooled data were also analyzed from the perspective of Hogarth's equation. The participants' average validities (the average correlation between a participant's answers and the correct answers), the group's average degrees of interpersonal correlation, the validity of the mean answer for each network (the correlation between the mean answers and the correct answers), and the prediction of Hogarth's equation of the validity of the mean answer were calculated for all three groups. The calculations of these quantities were done in the same way as the calculations of error and diversity above. Thus, the validity of the mean reported here is the mean of the corresponding score for each of the four groups. The average validity corresponds to the mean of the four means for each of the four groups (encompassing 7, 9, 9 and 13 correlations respectively). Finally, the average degree of interpersonal correlation was calculated by taking the mean of four means of the pairwise correlations in each group (encompassing 21, 36, 36, and 78 correlations respectively). The results can be seen in Table 5.

[5] We did not include error-bars in Figs. 3–5 for the following reasons: In the case of collective error, they do not provide meaningful insight into group level performance. Given that the collective error represents a mean of only four groups, the standard deviation seems meaningful only if one were to also include the variability across questions. In this case, though, the error bars would simply reflect the differences in question difficulty across a set of questions that were intentionally chosen to vary in difficulty. By contrast, average individual error—shown in Fig. 4—is a quantity that exhibits variation even within a question. However, that variability is explicitly represented by the diversity plot of 5, thus making error bars redundant.

[6] *p* values for individual error comparing round 2 to round 4 for the small world network: .01; for collective error: .03. The small-world network continues to show sporadic improvement beyond round 3 but there are no clear trends like those for rounds 1 and 2.

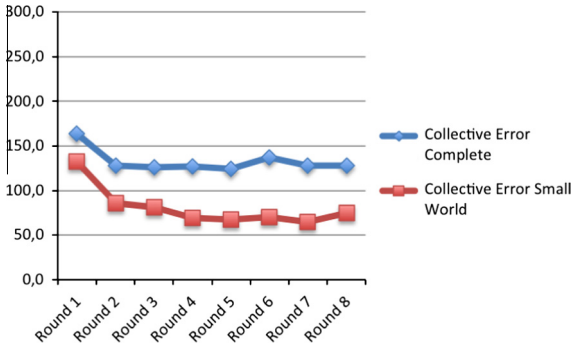

Fig. 3. Experiment 2: Collective error trends. Displayed are, by round and network, the error of the mean across all participants ($n = 38$), on a given question, averaged across all 10 questions.

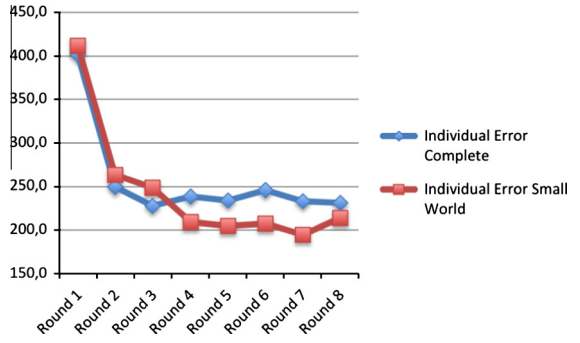

Fig. 4. Experiment 2: Individual error trends. Displayed are the mean individual errors, averaged across participants ($n = 38$) and question ($n = 10$), for each round and network type.

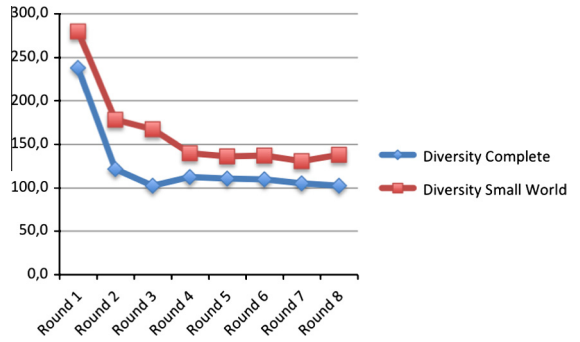

Fig. 5. Experiment 2: Diversity trends. Displayed is the mean variance in participants' answers, averaged across 10 questions, for each round and network type.

Across the four groups, and across the two network types, Hogarth's Eq. (2) predicted the validity of the mean answer with a high degree of accuracy. The overall degree of correlation between the validity of the mean answer and the predicted validity was very high (≈.93). The accuracy of the equation was comparable across the two network types. As is entailed by the equation, the mean validity outperformed the average individual validity across rounds and network-structures. The trends for the three relevant quantities can be seen in Figs. 6–8. As can be seen there is a stable reduction in collective error only for the small world network condition.

Table 5
Experiment 2: Correlations; means.

Network type	Rnd	Avg. val.	Avg. int. corr.	Val. of M. A	Hog. pred.
Complete	1	.765	.690	.898	.921
Complete	2	.850	.832	.921	.932
Complete	3	.855	.848	.918	.929
Complete	4	.863	.858	.924	.932
Complete	5	.867	.864	.924	.933
Complete	6	.863	.871	.916	.925
Complete	7	.868	.873	.921	.929
Complete	8	.866	.863	.923	.932
Small World	1	.764	.640	.917	.955
Small World	2	.843	.758	.942	.967
Small World	3	.852	.769	.946	.971
Small World	4	.878	.808	.951	.977
Small World	5	.881	.815	.951	.976
Small World	6	.879	.815	.950	.974
Small World	7	.884	.822	.952	.974
Small World	8	.877	.815	.947	.972

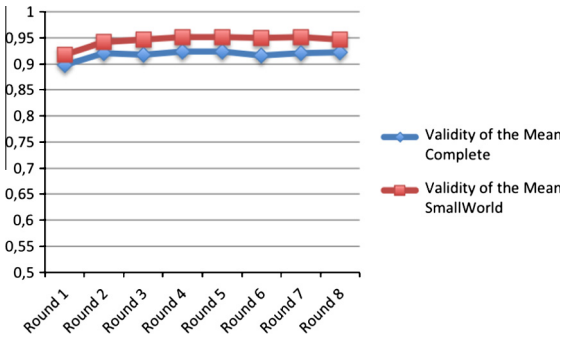

Fig. 6. Validity of the mean answer. Displayed for each network type, and across rounds, is the average, across the four groups, of the correlation between the set of true answers and the mean estimates of the participants in a group on each of the 10 questions.

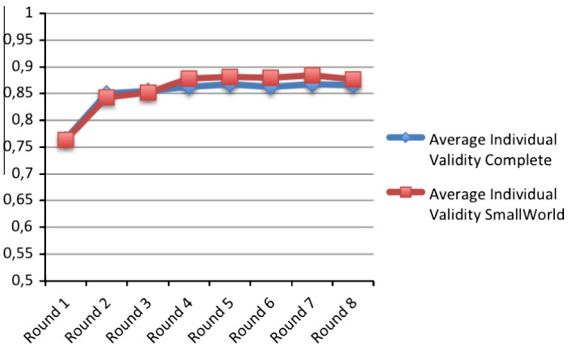

Fig. 7. Average validities. Displayed for each network type, and across rounds, is the mean across the four groups of the average individual correlation between participant estimates in that group and true answers.

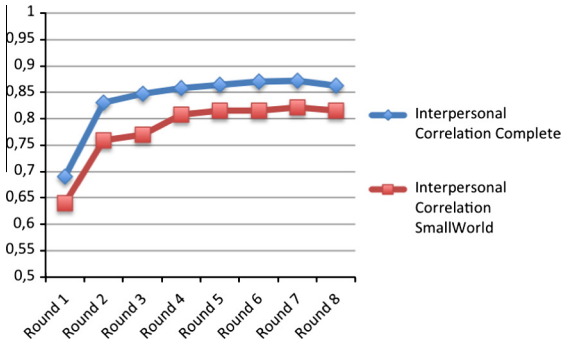

Fig. 8. Interpersonal correlation. Displayed, for each network type and across rounds, is the mean of the pairwise correlations between participants within each of the four groups-averaged across all four groups.

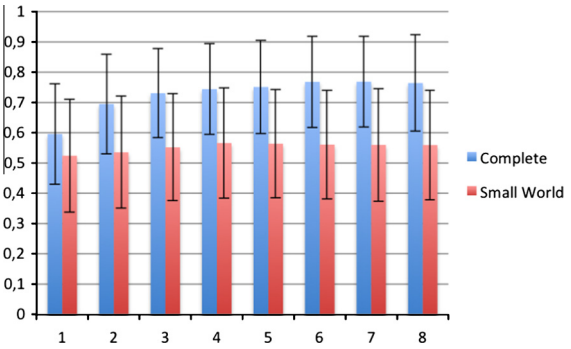

Fig. 9. Experiment 2: Excess correlation trends. Displayed are the means of the normalized excess correlation between participants within each of the four groups, averaged across groups. Error bars correspond to ±1 SD.

The data were also analyzed from the perspective of our measure of excess correlation, and the result can be seen in Fig. 9. The complete network increased more in excess correlation (+.1691) than did the small world network (+.0353). A two-tailed *t*-test on the final round revealed a highly significant difference between the excess correlation in the complete network and the small world network, $t = 9.75(169), p < .000001$.

11. Discussion

Experiment 2 replicates the results from Experiment 1 with respect to diversity and average individual error. Both quantities were significantly lowered in every round following the first one. Moreover, the small-world network, but not the complete network, showed a significant decrease in collective error for every round after the second one. This confirms the idea suggested by the Diversity Prediction Theorem, that by restricting information flow in a group, and thus preventing excessive decreases in diversity, one can generate a group structure such that both the average individual error and the group error is significantly reduced. The mechanism behind this, and the reason that the complete network did not decrease collective error to the same degree as the small world network, can be seen at work in Figs. 3–5. From these it can be seen that even though the individual error is comparable between the complete and the small-world network (Fig. 4), the diversity drops much lower in the complete network than in the small-world network (Fig. 5). The net effect of this, as entailed by the Diversity Prediction Theorem, is a bigger drop in collective error in the small-world network than in the complete network (Fig. 3).

The experiment thus constitutes an even stronger rebuttal of the seeming scepticism of Lorenz et al. (2011) concerning communication than did the first experiment, by showing that social interaction is consistent with significant decreases in both the average individual error *and* collective error.

From a correlational perspective, Experiment 2 also replicates Experiment 1 by again demonstrating the predictive accuracy of Hogarth's equation. Moreover, our measure of excess correlation now reveals an even bigger gap between the complete network and the small-world network than in Experiment 1. This confirms the results of Exp. 1 concerning the difference between the two network types. Given the very small increase in excess correlation (less than 4%) for the small world network, the measure of excess correlation indicates that, for this network structure, virtually all of the increase in correlation between participants (and the corresponding reduction in diversity) occurs as a mathematically necessary by-product of the increase in individual accuracy. By contrast, a considerable proportion of the inter-correlation that arises within the complete network is unrelated to the true answers to the questions.

In short, even though participants have more information to draw on in the complete network condition, this, if anything, hurts rather than helps the accuracy of their judgments. Note that this is not because participants cannot or do not incorporate more than a handful of pieces of evidence (cf. Yaniv & Milyavsky, 2007). That participants are not ignoring the additional information from others in the complete network is apparent from the very fact that their judgments become correlated to a greater extent in the complete network condition than in the small world network, even though these increases are not necessitated by individual accuracy. Participants are sensitive to the extra information available in the complete network condition, but that sensitivity, to a good extent, merely serves to amplify noise.

Finally, there is an interesting tension with respect to individual and collective error that arises from participants' specific use of others' information. As already observed in Exp. 1 above, participants in our study over-weight their own opinions, in keeping with previous findings in the literature on advice (e.g., Yaniv & Milyavsky, 2007), that is, they seemingly give more weight to their own estimate than they do to others. Participants in the complete network condition could have fairly easily (roughly) calculated the mean answer, and, on average, if they had adopted this answer, their individual accuracies would have been much higher (e.g., for round 2 this would have meant a drop in average individual error by 237.7 rather than 152.4). However, had they done so, they would

not have improved as a group at all, and missed out on the collective improvement that they did in fact obtain.[7] In other words, over-weighting of their own opinions led participants to less accurate individual responses than they could have otherwise obtained, but it is only due to that selective weighting that collective competence improved.

12. General discussion

We have provided what is, to the best of our knowledge, the first experimental study into the effects of network structure on the accuracy of group judgments in estimation tasks as they have long been a focus of interest, from Galton's classic (1907) study which prompted subsequent interest in the wisdom of crowds, through the extensive body of work on the impact of group deliberation on performance in social psychology between 1930 and 1970 (see e.g., Hill, 1982).

Specifically, we find clear evidence of effects of network structure on both collective and individual accuracy, whereby less densely connected groups outperform groups where every members' judgments are accessible to all. However, in all groups we find clear evidence against the claim (Lorenz et al., 2011) that access to others' judgments is detrimental to performance because of the reduction in diversity that it brings, as we found that individuals' average accuracy rose in response to information about others' estimates (supporting Farrell, 2011). Moreover, in the less densely connected networks even collective competence rose as a result of information exchange. Though collective competence (wisdom of the crowd) is necessarily a function of both individual competence and group diversity, less fully connected groups may increase individual accuracy sufficiently to offset the decrease in diversity information exchange brings about.

The need to add a further moderating factor—network structure—to the already complicated picture of when group influence is good or bad may, at first glance, seem worrying. However, we hope also to have shown that there is a wealth of formal tools, both from the literature on the aggregation of opinion and the literature on social networks, that may profitably be combined to make questions about group influence more tractable than they have been in the past.

Acknowledgements

The research reported in this article was funded by the Swedish Research Council through the framework project 'Knowledge in a Digital World' (Erik J. Olsson, PI) and the Swedish Research Council's Kerstin Hesselgren Professorship (Ulrike Hahn).

Appendix A. Questions

All questions pertain to the year 2011 unless otherwise stated.

A.1. Experiment 1

WU	What percent of the Swedes are 15–24 years? (13%)
S1	What percent of Sweden is covered by agricultural land? (7.6%)
S2	What percent of Swedish university students study a humanities subject? (15.4%)
S3	What percent of the Swedes have undergone higher education? (24%)
S4	What percent of the Swedes who entered the university in 2011 have at least one university educated parent? (35%)
S5	What percent of energy use in Sweden is renewable? (48%)
S6	What percent of the Swedes roam in the forest more than 5 times a year? (55.5%)
S7	What percent of Swedish men aged 35–44 are overweight? (61%)
S8	What percent of the Swedes between 15–74 have or are looking for work? (71%)
S9	What percent of the electorate Swedes voted in parliamentary elections in 2010? (84.6%)
S10	What percent of Swedish Internet users over 12 years have ever used Google? (97%)
R1	What percent of Sweden's surface is developed? (2.9%)
R2	What percent of the Swedes live in Skåne county? (13.2%)
R3	What percent of the deceased in Sweden died of tumors? (25%)
R4	What percent of crimes in Sweden are solved? (39%)
R5	What percent of newly admitted graduate students at Lund University are women? (44.4%)
R6	What percent of Sweden is covered by forest? (53.1%)
R7	What percent of the Swedes aged 16–84 went for a holiday for at least one week? (60%)
R8	What percent of the deceased in Sweden were over 75 years? (72%)
R9	What percent of Swedish children in their sixth school-year are vaccinated against polio? (96%)
R10	What percent of the electorate Swedes took part in the referendum on the introduction of the euro in 2003? (82.6%)
C1	What percent of Swedish women have Anna as their given name? (2.3%)
C2	What percent of those starting the university have a foreign background? (17%)
C3	What percent of Swedish adolescents aged 15–24 are unemployed? (22.9%)
C4	What percent of the Swedes are married? (33.9%)
C5	What percent of the vote in parliamentary elections in 2010 went to the Alliance? (49.3%)
C6	What percent of Swedish MPs are men? (55%)
C7	What percent of Swedish boys between 13–15 years engage in sports at least once a week? (62%)

[7] We thank Igor Volzhanin for this observation.

C8	What percent of Swedish girls born 1999–2001 have received at least one dose of the HPV-vaccine? (78%)
C9	What percent of the children who completed their ninth school-year passed Swedish, mathematics and English? (87.2%)
C10	What percent of Swedish girls aged 10–12 years have their own room? (92%)

A.2. Experiment 2

WU	What percent of the Swedes are 15–24 years? (13%)
1	What percent of Swedish women have Anna as their given name? (2.3%)
2	What percent of those starting the university have a foreign background? (17%)
3	What percent of the Swedes are married? (33.9%)
4	What percent of the vote in parliamentary elections in 2010 went to the Alliance? (49.3%)
5	What percent of Swedish MPs are men? (55%)
6	What percent of Swedish boys between 13–15 years engage in sports at least once a week? (62%)
7	What percent of Swedish girls born 1999–2001 have received at least one dose of the HPV-vaccine? (78%)
8	What percent of the children who completed their ninth school-year passed Swedish, mathematics and English? (87.2%)
9	What percent of the Swedes live in Skåne county? (13.2%)
10	What percent of the deceased in Sweden died of tumors? (25%)
11	What percent of newly admitted graduate students at Lund University are women? (44.4%)
12	What percent of Sweden is covered by forest? (53.1%)
13	What percent of the Swedes aged 16–84 went for a holiday for at least one week? (60%)
14	What percent of Swedish children in their sixth school-year are vaccinated against polio? (96%)
15	What percent of Sweden is covered by agricultural land? (7.6%)
16	What percent of energy use in Sweden is renewable? (48%)
17	What percent of the Swedes roam in the forest more than 5 times a year? (55.5%)
18	What percent of the Swedes between 15–74 have or are looking for work? (71%)
19	What percent of the electorate Swedes voted in parliamentary elections in 2010? (84.6%)
20	What percent of Swedish Internet users over 12 years have ever used Google? (97%)

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2015.04.013.

References

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America, 22*, 725–730.

Coady, C. A. J. (1992). *Testimony. A philosophical study*. Oxford University Press.

Doer, B., Fouz, M., & Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM, 55*(6), 70.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*(1), 158–172.

Erdos, P., & Rényi, I. (1959). On random graphs. I. *Publicationes Mathematicae, 6*, 290–297.

Farrell, S. (2011). Social influence benefits the wisdom of individuals in the crowd. *Proceedings of the National Academy of Sciences, 108*(36), E625.

Freeman, L. C., Roeder, D., & Mulholland, R. R. (1979). Centrality in social networks: II experimental results. *Social Networks, 2*, 119–141.

Friedkin, N. E., & Johnsen, E. C. (1999). Social influence networks and opinion change. *Advances in Group Processes, 16*(1), 1–29.

Galton, F. (1907). Vox populi. *Nature, 75*, 450–451.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. McGraw-Hill.

Gigone, D., & Hastie, R. (1997). Proper analysis of the accuracy of group judgments. *Psychological Bulletin, 121*(1), 149.

Goldstone, R. L., & Gureckis, T. M. (2009). Collective behavior. *Topics in Cognitive Science, 1*(3), 412–438.

Goldstone, R. L., Roberts, M. E., & Gureckis, T. M. (2008). Emergent processes in group behavior. *Current Directions in Psychological Science, 17*(1), 10–15.

Hertwig, R. (2012). Tapping into the wisdom of the crowd-with confidence. *Science, 336*(6079), 303–304.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science: A Journal of the American Psychological Society/APS, 20*(2), 231–237.

Hill, G. W. (1982). Group versus individual performance: Are N+ 1 heads better than one? *Psychological Bulletin, 91*(3), 517.

Hogarth, R. M. (1977). Methods for aggregating opinions. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 231–255). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behaviour and Human Performance, 21*, 40–46.

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Katz, N., Lazer, D., Arrow, H., & Contractor, N. (2004). Network theory and small groups. *Small Group Research, 35*(3), 307–332.

Kearns, M. (2006). An experimental study of the coloring problem on human subject networks. *Science, 313*(5788), 824–827.

King, A. J., Cheng, L., Starke, S. D., & Myatt, J. P. (2012). Is the true wisdom of crowd to copy successful individuals? *Biology Letters, 8*, 197–200.

Knight, H. C. (1921). *A comparison of the reliability of group and individual judgments*. Master's thesis, Columbia University.

Kretzschmar, M., & Morris, M. (1996). Measures of concurrency in networks and the spread of infectious disease. *Mathematical Biosciences, 133*(2), 165–195.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly, 52*, 667–694.

Leavitt, H. J. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology, 46*(1), 38.

Levine, J. M., & Moreland, R. L. (2012). A history of small group research. *Handbook of the History of Social Psychology*, 383.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025.

Lorge, I., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance: 1920–1957. *Psychological Bulletin, 55*, 337–372.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*, 71–87.
Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General, 137*(3), 422–433.
McGrath, J. E., Arrow, H., & Berdahl, J. L. (2000). The study of groups: Past, present, and future. *Personality and Social Psychology Review, 4*(1), 95–105.
Page, S. E. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, New Jersey: Princeton University Press.
Rauhut, H., & Lorenz, J. (2010). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology, 55*(191–197).
Shaw, M. E. (1932). Comparison of individuals and small groups in the rational solution of complex problems. *American Journal of Psychology, 44*(3), 491–504.
Shaw, M. E. (1964). Communication networks. In L. Berkowitz (Ed.), *Advances in experimental psychology* (pp. 111–147). New York: Academic Press.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Brown.
Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647.
Wallsten, T., Budescu, D., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10*, 243–268.
Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology, 105*(2), 493–527.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.
Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes, 103*(1), 104–120.